# ESWC Workshop report: Empirical 2014

http://2014.empirical-ws.org/

Co-chairs: Kjetil Kjernsmo, Maria-Esther Vidal, Jacco van Ossenbruggen

0 submissions
24 people present

**Keynote**
"Evaluation in Semantic Web Research" by Abraham Bernstein
Previous version of the slide deck:
http://www.merlin.uzh.ch/publication/show/8514

For more info, recommended reading includes:
“Is This Really Science? The Semantic Webber’s Guide to Evaluating Research Contributions” by Abraham Bernstein & Natasha Noy
https://www.merlin.uzh.ch/contributionDocument/download/6915

**Invited Position Paper Talks:**
“*The role of annotation in reproducibility*” By Oscar Corcho
"*An Experience on Empirical Research about RDF Stream Processing*" By Daniele Dell'Agilio

## Trial

Judge: Frank van Harmelen
Prosecutor: Jacco van Ossenbruggen
Witnesses for the prosecution: Kjetil Kjernsmo, Miel Vander Sande
Defense attorney: Maria-Esther Vidal
Defense witnesses: Jérôme Euzenat, Oscar Corcho

Benchmarking was being accused of two charges,

1. Not contributing in the advancement of our field
2. To be detrimental in blocking other methods to do so

Kjernsmo and Vander Sande testified supporting these accusations, after which the defense argued in favor for *real* benchmarks over mere “proto” benchmarks. The audience got the chance to cross-examine the witnesses after which judge Van Harmelen ruled fairly and wisely.

The defendant was cleared on the first charge by lack of convincing evidence by the prosecution and evidence to the contrary by OAEI witness Euzenat. Benchmarking was found guilty on the second charge. Considering the lack of bad intent and the immature age, the defendant was released on probation, with a retrial scheduled for next year.

## Breakout groups

Groups were formed on the topics of:

**Checklists**:

Can we develop checklists analogous to those in use in other disciplines, for semantic web research? Group produced an initial stab at a checklist for emperical SW paper, reproducable SW research and a checklist for negative result papers

**Benchmarking the benchmarks**

If there are good and bad benchmarks, how can we identify the good ones, how can we improve existing ones and develop new? Addressing reproducibility, representativeness, overfitting, community bias, etc.

## Dagstuhl seminar

Is this a good topic for a future Dagstuhl seminar? Group discussed the scope of the topic, that it was also relevant for other benchmark driven disciplines, including ML, NLP, IR, DB